

УДК 303.224.74

Д. А. Федерякин

dafederiakina@hse.ru

Национальный исследовательский университет «Высшая школа экономики»,
Москва, Россия**КЛАССИФИКАЦИЯ БИФАКТОРНЫХ МОДЕЛЕЙ ***

Бифакторные модели отражают современный запрос на композитные инструменты измерения – те, которые позволяют сообщать пользователю как балл по всему тесту, так и баллы по частям теста. Эта работа предлагает классификацию бифакторных моделей современной теории тестирования. Классификация основана на четырех критериях: 1) количество параметров на задание; 2) функция связи параметров с вероятностью ответа; 3) структура порогов в полиномических заданиях; 4) косоугольность размерностей респондентов.

Ключевые слова: современная теория тестирования, бифакторные модели, классификация бифакторных моделей, косоугольные бифакторные модели.

Denis A. Federiakina

dafederiakina@hse.ru

National Research University Higher School of Economics, Moscow, Russia

CLASSIFICATION OF BIFACTOR MODELS

Bifactor models reflect the current popularity of composite tests, which allow for a report of the general and specific test scores. This paper proposes a classification of bifactor models from the item response theory paradigm. The classification is based on four criteria: (1) the number of parameters per item, (2) link function, (3) structure of thresholds in polytomous items, (4) obliqueness of persons' dimensions.

Keywords: item response theory, bifactor models, classification of bifactor models, oblique bifactor models.

Бифакторные модели современной теории тестирования (Item Response Theory, IRT) являются одним из наиболее популярных методологических решений для психометрического моделирования с самого момента их разработки [1]. Эти модели позволяют одновременно извлечь из данных некоторый общий фактор, который проявляется во всех заданиях, а также специфические факторы, которые проявляются только в части теста. Их популярность сейчас определяется популярностью композитных инструментов измерения – тех, которые призваны выдавать

© Федерякин Д. А., 2021

* Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-29-14110 «Использование контекстной информации и информации из цифровой среды оценивания при измерении индивидуального прогресса учащихся начальной школы с помощью цифровых технологий».

тестовый балл как по всему тесту, так и по его компонентам. В таких случаях применение одномерных моделей IRT приводит к невозможности доверять оценкам параметров, потому что их допущение о локальной независимости заданий на параметрах респондентов не выполняется. Традиционные многомерные модели же, лишены возможности извлекать общий фактор – композитную характеристику респондентов, в которой, как правило, заинтересованы исследователи и пользователи результатов. Подробнее, психометрическое моделирование таких инструментов описано в [2–4].

Целью данной работы является создание классификации бифакторных моделей, которая позволит исследователям делать осознанный выбор модели (и дизайна инструмента измерения), а также подсветит направления дальнейших методологических исследований в этой области. В качестве оснований для классификации мы используем четыре критерия: 1) количество параметров на задание; 2) функция связи параметров с вероятностью ответа; 3) структура порогов в политомических заданиях; 4) косоугольность размерностей респондентов.

1. Количество параметров на задание. В качестве первого критерия мы используем традиционное разделение дихотомических IRT-моделей по количеству параметров на задание. Согласно этому критерию, IRT-модели могут делиться на однопараметрические (One Parameter Logistic Model (1PL), где дискриминативность одинакова для всех заданий [5]), двухпараметрические (2PL, где задания имеют свободный параметр дискриминативности [6]), и трехпараметрические (3PL, где задания имеют свободный параметр вероятности случайного угадывания вдобавок к дискриминативности [6]). Несмотря на то, что в литературе описаны попытки введения в модель дополнительных параметров (например, вероятности случайной ошибки [7], или параметр немонотонности характеристической кривой задания (unfolding parameter) [8]), в этой работе мы ограничиваемся описанием только первых двух моделей. Это связано с тем, что 3PL модель не идентифицирована, т. е. не имеет глобального максимума правдоподобия (как и её обобщения [9]). Как следствие, вероятность угадывания может быть введена в модель только в случае, если дискриминативность заданий не оценивается, создавая не 3PL, а 1PL-G (1PL-Guessing) модель [10].

1PL дихотомическая бифакторная модель в скалярном виде задается уравнением:

$$\text{logit}\left(P\left(X_{ip}=1\right)\right)=\theta_{gp}+\theta_{s(d)p}-\xi_i, \quad (1)$$

где $\text{logit}\left(P\left(X_{ip}=1\right)\right)$ – это логарифмический шанс того, что ответ респондента p на задание i будет равен 1; θ_{gp} – это значение общего фактора

у респондента p ; $\theta_{s(d)p}$ – это значение специфического фактора d у респондента p ; ξ_i – это трудность задания i .

В уравнение 2PL дихотомической бифакторной модели добавляется параметр дискриминативности заданий:

$$\text{logit}\left(P\left(X_{ip}=1\right)\right)=\alpha_{gi}\theta_{gp}+\alpha_{s(d)i}\theta_{s(d)p}-\xi_i, \quad (2)$$

где α_{gi} – дискриминативность задания i к общему фактору; $\alpha_{s(d)i}$ – дискриминативность задания i к специфическому фактору d .

Крайне популярен специальный случай модели (2), когда внутри каждого специфического фактора выдерживается условие $\frac{\alpha_{gi}}{\alpha_{s(d)i}} = \frac{\alpha_{gj}}{\alpha_{s(d)j}}$ для всех $i \neq j$. Такая бифакторная модель является Шмид – Лейман трансформированной моделью с фактором высокого порядка [11]:

$$\text{logit}\left(P\left(X_{ip}=1\right)\right)=\alpha_i\left(\theta_{gp}+\alpha_{s(d)}\theta_{s(d)p}\right)-\xi_i, \quad (3)$$

где α_i – дискриминативность задания i ; $\alpha_{s(d)}$ – нагрузка специфического фактора d на общий.

Сравнение уравнений (2) и (3) показывает, что менее вероятно, что модель с фактором высокого порядка подойдет данным лучше, чем бифакторная, потому что она имеет меньше параметров [12]. Кроме того, из этого следует, что трансформация Шмида – Леймана позволяет оценить модель с фактором высокого порядка как бифакторную модель в случаях, когда, например, тест состоит из двух специфических факторов. Также это подразумевает, что в традиции 1PL-моделирования, отношения между этими моделями обратны и 1PL бифакторная модель является частным случаем 1PL модели с фактором второго порядка, которая может быть описана как:

$$\text{logit}\left(P\left(X_{ip}=1\right)\right)=\theta_{gp}+\alpha_{s(d)}\theta_{s(d)p}-\xi_i. \quad (4)$$

Уравнение 1PL-G дихотомической бифакторной модели выглядит как:

$$P\left(X_{ip}=1\right)=c_i+(1-c_i) \cdot \text{logit}^{-1}\left(\theta_{gp}+\theta_{s(d)p}-\xi_i\right), \quad (5)$$

где c_i – это вероятность случайно угадать ответ на задание i .

2. Функция связи параметров с вероятностью ответа. В случае использования политомических данных, исследователь, обычно, выбирает между функциями логит-связи, которые моделируют или поочередную

вероятность перехода в более высокую категорию ответа *при условии, что респондент уже находится в предыдущей категории* (adjacent logit, сопредельная логит-связь), или вероятность того, что респондент наберет балл в данной категории или выше при его восхождении по категориям ответа (cumulative logit, кумулятивная логит-связь). В случае дихотомических данных, различия между этими типами связи нет. Для полиномической вариации уравнения (2) модель сопредельной логит-связи выглядит как:

$$\ln \left(\frac{P(X_{ip} = x)}{P(X_{ip} = x-1)} \right) = \alpha_{gi} \theta_{gp} + \alpha_{s(d)i} \theta_{s(d)p} - \xi_{ik}, \quad (6)$$

где $P(X_{ip} = x)$ – вероятность ответа в категории x ; ξ_{ik} – порог между категориями $x-1$ и x . Для идентификации, самая первая (минимальная) категория в каждом задании рассматривается как референтная (за нее начисляется балл 0).

Уравнение (6) описывает бифакторную обобщенную модель частичного оценивания (Generalized Partial Credit Model, GPCM [13]) – обобщение 2PL модели на сопредельную связь. Её частным случаем является модель частичного оценивания (Partial Credit Model, PCM [14]) в случае, когда все дискриминативности заданий равны. Уравнение (7) описывает бифакторную модель градуированного ответа (Graded Response Model, GRM [15]) – обобщение 2PL модели на кумулятивную связь:

$$\ln \left(\frac{P(X_{ip} \geq x)}{P(X_{ip} < x)} \right) = \alpha_{gi} \theta_{gp} + \alpha_{s(d)i} \theta_{s(d)p} - \xi_{ik}. \quad (7)$$

Интерпретационно выбор между этими моделями сводится к представлению исследователя о процессе ответа на задание. Если исследователь подразумевает, что респондент «шаг за шагом» восходит внутри каждого задания от самой низкой категории к более высокой до тех пор, пока его способности хватает на то, чтобы сделать переход, то исследователю следует выбирать модели кумулятивной связи. Если исследователь считает, что респондент по ходу ответа сразу «оказывается» в наблюдаемой категории ответа, то ему следует выбирать модели сопредельной связи.

Сравнение уравнений (6) и (7) подчеркивает и разницу в возможностях, которую предоставляют эти модели. В моделях сопредельной связи может возникнуть ситуация неупорядоченности порогов категорий, потому что, например, респонденты «перепрыгивают» одну из категорий [16]. В моделях кумулятивной связи же пороги категорий всегда будут упорядочены. Несмотря на то, что оба семейства дают достаточно высокую корреляцию способностей респондентов (около 0,9), она

не равна единице – следовательно, они ранжируют респондентов по разным способностям, которые различаются в предполагаемом процессе ответа. Важно также отметить, что благодаря полному разделению параметров респондентов и заданий (в 1PL вариациях), модели сопредельной связи позволяют конструировать комплексную композитность заданий. Она заключается в том, что переход из разных категорий в более высокие может зависеть от разного набора способностей респондентов внутри одного и того же задания [17].

Интересно отметить, что ограниченная версия GRM (Constrained GRM, C-GRM) на случай с одинаковыми дискриминативностями не описана в литературе, поскольку она лишена практического смысла. Исследователи, занимающиеся развитием моделей с одинаковыми дискриминативностями, обычно сосредоточены на получении простого разделения параметров заданий и респондентов, что естественным образом приводит их к моделям сопредельной связи. В свою очередь же, C-GRM приводит к ухудшению согласия модели с данными по сравнению с GRM, не обеспечивая простого алгебраического разделения параметров [18].

3. Структура порогов в политомических заданиях. Политомические модели могут различаться по структуре порогов в заданиях. Рассмотрим частный случай РСМ – модель рейтинговой шкалы (Rating Scale Model, RSM [19, 20]):

$$\ln \left(\frac{P(X_{ip} = x)}{P(X_{ip} = x-1)} \right) = \theta_{gp} + \theta_{s(d)p} - \tau_k - \delta_i, \quad (8)$$

где τ_k – порог между категориями $x-1$ и x (для всех заданий); δ_i – трудность задания i .

Декомпозиция ξ_{ik} на τ_k и δ_i позволяет ввести в модель одинаковую структуру порогов у всех заданий, отражающую теоретическое допущение об одинаковости интерпретации соответствующих ответных категорий в разных заданиях. Это допущение осмысленно в случаях психологических опросников со шкалой Ликерта, когда формулировка всех ответных категорий одинакова для всех заданий.

Существует специальная версия GRM – модифицированная модель градуированного ответа (Modified Graded Response Model, M-GRM [21]), которая описывает такую же декомпозицию, но в моделях куммулятивной связи. Однако она, как и обобщенная модель рейтинговой шкалы (Generalized Rating Scale Model, G-RSM [22] – расширение RSM на случай разных дискриминативностей заданий), лишена практического смысла. Даже несмотря на то, что структура порогов во всех заданиях в таких моделях одинаковая, различная дискриминативность приводит к различиям характеристических кривых категорий разных заданий.

А именно эта характеристика является отражением допущения об одинаковом смысле категорий ответа.

Если опросник состоит из разных ответных шкал (например, одна шкала сформулирована в терминах частоты, а другая – степени согласия с утверждением), то это должно быть отражено в модели [23]:

$$\ln \left(\frac{P(X_{ip} = x)}{P(X_{ip} = x-1)} \right) = \theta_{gp} + \theta_{s(d)p} - \tau_{k(r)} - \delta_i, \quad (9)$$

где $\tau_{k(r)}$ – порог между категориями $x-1$ и x в ответной шкале r .

4. Корреляции между размерностями респондентов.

Традиционные бифакторные модели являются ортогональными, т. е. требуют от исследователя сделать допущение об отсутствии корреляций между всеми размерностями респондентов для идентификации модели. Из-за этого невозможно интерпретировать специфические факторы как компоненты общего фактора.

В ответ на это ограничение был предложен ряд косоугольных бифакторных моделей, которые позволяют преодолеть его. Так, расширенная Раш-тестлет модель (Extended Rasch Testlet Model, ETM [24]) позволяет оценить корреляции специфических факторов с общим, но сохраняет ортогональность специфических факторов друг к другу. Другой группой моделей является субшкальная модель (Subdimensional Model, SM [25]) и обобщенная субшкальная модель (Generalized Subdimensional Model, GSM; [26]), которые позволяют оценить корреляции специфических факторов друг с другом, но сохраняют их ортогональность к общему фактору. Различия между этими моделями заключаются в том, что в то время, как SM допускает, что дисперсии всех специфических факторов равны (что ухудшает согласие модели с данными), GSM не делает такого допущения. Наконец, полностью косоугольная бифакторная Раш-модель (Completely Oblique Rasch Bifactor, CORB model), позволяет оценить все корреляции между всеми размерностями латентного пространства параметров респондентов [27].

Federiakin & Wilson [28] показали, что CORB модель идентифицирована в случае специфических структур размерности теста, например, когда есть хотя бы одно задание, которое не нагружает ни один специфический фактор (т. е. фактически задает общий). Однако множество структур размерности теста, которые идентифицируют CORB модель, является бесконечно большим, что приводит к необходимости проведения процедуры, описанной в [29] для каждого теста, чтобы понять, идентифицирована ли CORB модель в данном случае. Эта процедура позволяет проанализировать, идентифицирована ли желаемая Раш-модель для данной многомерной структуры теста.

При этом ETM является частным случаем CORB (а обычная ортогональная бифакторная Раш-модель является частным случаем ETM),

однако GSM – нет. Ограничения, введенные в матрицу ковариаций параметров респондентов в ETM (и ортогональную бифакторную модель), являются элементарными ограничениями этой матрицы из CORB. В GSM же один из специфических факторов задан как отрицательная сумма остальных специфических факторов. Однако CORB позволяет вывести (как частный случай) субшкальную косоугольную бифакторную Раш-модель (Subdimensional Oblique Rasch Bifactor, SORB model), вводя прямые ограничения в оцениваемую матрицу ковариаций параметров респондентов, похожие на ограничения ETM, но отражающие структуру GSM. Возможность оценивать корреляции между различными факторами в косоугольных бифакторных моделях позволяет использовать каждый из факторов как коллатеральную информацию о респондентах для всех остальных факторов – информацию, которая не меняет интерпретацию параметров модели, но повышает надежность измерения [30].

Все описанные косоугольные бифакторные модели были предложены и изучены в парадигме 1PL моделей. Их расширение на 2PL должно быть изучено дальше. Тем не менее в 2PL существует описание косоугольных бифакторных моделей, которые идентифицированы за счет того, что сумма ковариаций всех факторов равна нулю, или сумма ковариаций и дисперсий всех факторов равна нулю [17]. Однако интерпретация параметров этих моделей невозможна, поскольку невозможно иметь такие ожидания от данных. Более того, поведение этих моделей не было изучено – на сегодняшний день, только теоретически описана их структура. Другой возможностью косоугольного 2PL моделирования является применение косоугольных бифакторных моделей из эксплораторного факторного анализа [31], что обеспечивает наиболее полную репрезентацию матрицы заданий в пространстве латентных факторов.

Заключение. Данная работа предлагает классификацию бифакторных моделей по четырем критериям: 1) количество параметров на задание; 2) функция связи параметров с вероятностью ответа; 3) структура порогов в политомических заданиях; 4) косоугольность размерностей респондентов. Описание такой классификации позволяет исследователям делать более осознанный выбор при определении методологии обработки результатов теста. Кроме того, эта классификация позволяет определить дальнейшие перспективные направления исследований в области бифакторного моделирования. Однако она не лишена ограничений. В частности, она описывает характеристики, присущие IRT-моделям в целом (например, критерий 2), которые тем не менее важны в контексте бифакторных моделей, потому что косоугольные бифакторные модели описаны в политомическом случае для сопредельной логит-связи. Возможности косоугольного бифакторного моделирования в случае кумулятивной логит-связи должны быть изучены в дальнейшем. Кроме того, эта классификация не затрагивает бифакторные модели из факторного анализа, поскольку они эквивалентны моделям с кумулятивной логит-связью и нормализующей константой [32]

или кумулятивной пробит-связью, при использовании полихорических корреляций для построения матрицы корреляций заданий [33]. Также данная классификация не затрагивает различные способы параметризации модели [34].

Список литературы

1. Holzinger K. J., Swineford F. The bi-factor method // *Psychometrika*. 1937. № 2(1). P. 41–54.
2. Федерякин Д. А., Карданова Е. Ю. Психометрическое моделирование сложных конструкторов // *Информатизация образования и методика электронного обучения: цифровые технологии в образовании* / под ред. М. В. Носков. Красноярск, 2020. Т. 2. С. 242–248.
3. Federiakin D., Kardanova E. Psychometrical Modeling of Components of Composite Constructs: Recycling Data Can Be Useful // *Informatization of Education and E-learning Methodology: Digital Technologies in Education* / М. Noskov, А. Semenov, S. Grigoriev (Eds.). Krasnoyarsk, 2020. P. 157–171.
4. Federiakin D., Larina G., Kardanova E. Measuring Basic Mathematical Literacy in Elementary School // *Educational Studies*. 2021. №2. P. 199–226.
5. Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago: MESA, 1993. 199 p.
6. Birnbaum A. L. Some latent trait models and their use in inferring an examinee's ability // *Statistical theories of mental test scores* / F.M. Lord, M.R. Novick (Eds.). Charlotte, 1968. P. 397–479.
7. Barton M. A., Lord F. M. An upper asymptote for the three-parameter logistic item-response model. Princeton, NJ: Educational Testing Service, 1981. 24 p.
8. Andrich D., Luo G. A hyperbolic cosine latent trait model for unfolding dichotomous single-stimulus responses // *Applied Psychological Measurement*. 1993. №17(3). P. 253–276.
9. San Martín E., González J., Tuerlinckx F. On the unidentifiability of the fixed-effects 3PL model // *Psychometrika*. 2015. № 80 (2). P. 450–467.
10. Martín E. S., Del Pino G., de Boeck P. IRT models for ability-based guessing // *Applied Psychological Measurement*. 2006. № 30 (3). P. 183–203.
11. Schmid J., Leiman J. M. The development of hierarchical factor solutions // *Psychometrika*. 1957. № 22(1). P. 53–61.
12. Mansolf M., Reise S.P. When and why the second-order and bifactor models are distinguishable // *Intelligence*. 2017. №61. P. 120–129.
13. Muraki E. A generalized partial credit model: Application of an EM algorithm // *ETS Research Report Series*. 1992. № 1. P. 1–30.
14. Masters G. N. A Rasch model for partial credit scoring // *Psychometrika*. 1982. № 47(2). P. 149–174.
15. Samejima F. Estimation of latent ability using a response pattern of graded scores // *Psychometrika*. 1969. № 34 (2), P. 1–97.
16. Adams R. J., Wu M. L., Wilson M. The Rasch rating model and the disordered threshold controversy // *Educational and Psychological Measurement*. 2012. № 72 (4). P. 547–573.
17. Robitzsch A., Kiefer T., Wu M. Package ‘TAM’: Test Analysis Modules. v. 3.7-16. 2021. CRAN Team. 225 p.
18. Wright B. D., Masters G. N. Rating scale analysis. Chicago: MESA, 1982. 223 p.
19. Andrich D. A rating formulation for ordered response categories // *Psychometrika*. 1978. №43. P. 561–573.

20. Andrich D. Application of a psychometric rating model to ordered categories which are scored with successive integers // *Applied Psychological Measurement*. 1978. № 2. P. 581–594.
21. Muraki E. Fitting a polytomous item response model to Likert-type data // *Applied Psychological Measurement*. 1990. № 14(1). P. 59–71.
22. Wang W. C., Wu S. L. The random-effect generalized rating scale model // *Journal of Educational Measurement*. 2011. № 48 (4). P. 441–456.
23. Kanonire T., Federiakin D. A., Uglanova I. L. Multicomponent framework for students' subjective well-being in elementary school // *School Psychology*. 2020. № 35 (5). P. 321–331.
24. Random Parameter Structure and the Testlet Model: Extension of the Rasch Testlet Model. / I. Paek [et al.] // *Journal of Applied Measurement*. 2009. № 10. P. 394–407.
25. Brandt S. Estimation of a Rasch model including subdimensions // *IERI Monograph Series. Issues and Methodologies in Large-Scale Assessments* / von Davier M., Hastedt D. (Eds.). Princeton, NJ, 2008. P. 51–70.
26. Brandt S. Concurrent unidimensional and multidimensional calibration within item response theory // *Pensamiento Educativo. Revista de Investigación Educativa Latinoamericana*. 2017. №54. P. 1–16.
27. Federiakin D. On the Critique of the Oblique Bifactor Models: Conditions for Identification and Problems in Interpretation // Paper presented at the International Meeting of the Psychometric Society, 2021.
28. Federiakin D., Wilson M.R. The Completely Oblique Rasch Bifactor Model // *Journal of Applied Measurement*. In press.
29. Volodin N., Adams R.J. The Estimation of Polytomous Item Response Models with Many Dimensions // *ACER*. 2002. № 12. P. 1–34.
30. Федерякин Д. А., Углонова И. Л., Скрыбин М. А. Новые источники информации в компьютерном тестировании // *Вестник Томского государственного университета* (в печати).
31. Jennrich R. I., Bentler P. M. Exploratory bi-factor analysis: The oblique case // *Psychometrika*. 2012. № 77 (3). P. 442–454.
32. Wirth R.J., Edwards M.C. Item factor analysis: current approaches and future directions // *Psychological methods*. 2007. № 12 (1). P. 58–79.
33. Camilli G. The scaling constant D in item response theory // *Open Journal of Statistics*. 2017. №7(5). P. 780–785.
34. Estimation of an IRT model by Mplus for dichotomously scored responses under different estimation methods / I. Paek [et al.] // *Educational and Psychological Measurement*. 2018. № 78 (4). P. 569–588.